Network theory to understand microarray studies of complex diseases

Benson, M.; Breitling, R.

Published in:
Current Molecular Medicine

DOI:
10.2174/156652406778195044

Network Theory to Understand Microarray Studies of Complex Diseases

Mikael Benson*,[1] and Rainer Breitling[2]

[1]*Department of Pediatrics, Queen Silvia Children's Hospital, SE-41685 Gothenburg, Sweden*

[2]*Groningen Bioinformatics Centre, University of Groningen, 9751 NN Haren, The Netherlands*

Abstract: Complex diseases, such as allergy, diabetes and obesity depend on altered interactions between multiple genes, rather than changes in a single causal gene. DNA microarray studies of a complex disease often implicate hundreds of genes in the pathogenesis. This indicates that many different mechanisms and pathways are involved. How can we understand such complexity? How can hypotheses be formulated and tested?

One approach is to organize the data in network models and to analyze these in a top-down manner. Globally, networks in nature are often characterized by a small number of highly connected nodes, while the majority of nodes have few connections. The highly connected nodes serve as hubs that affect many other nodes. Such hubs have key roles in the network. In yeast cells, for example, deletion of highly connected proteins is associated with increased lethality, compared to deletion of less connected proteins.

This suggests the biological relevance of networks. Moving down in the network structure, there may be sub-networks or modules with specific functions. These modules may be further dissected to analyze individual nodes. In the context of DNA microarray studies of complex diseases, gene-interaction networks may contain modules of co-regulated or interacting genes that have distinct biological functions. Such modules may be linked to specific gene polymorphisms, transcription factors, cellular functions and disease mechanisms. Genes that are reliably active only in the context of their modules can be considered markers for the activity of the modules and may thus be promising candidates for biomarkers or therapeutic targets.

This review aims to give an introduction to network theory and how it can be applied to microarray studies of complex diseases.

INTRODUCTION

Complex diseases involve multiple genes. In allergic disease, for example, genome-wide linkage studies have shown more than 20 susceptibility loci. DNA microarray studies reveal hundreds of differentially expressed genes in cells and tissues from patients with complex diseases [1]. In a typical microarray article these genes are listed, some commented on, and one or two chosen for detailed analysis. In essence, such an article is based on two scales, one global and one detailed. While the detailed scale can give new and interesting information, the reason for highlighting a specific gene, and its importance relative to other genes, are often quite subjective. Moreover, analysis on the global level may be unstructured and uninformative. A structured way of describing both the global and detailed structure of the results is desirable. Essentially, this is the same idea as with a microscopic examination. You start with a low magnification to get a general overview of the tissue and then increase magnification to examine selected substructures and single cells. In this article the possibility of using network approaches for such top-down analysis of microarray data will be discussed.

*Address correspondence to this author at the Department of Pediatrics, Queen Silvia Children's Hospital, SE-41685 Gothenburg, Sweden; E-mail: mikael.benson@vgregion.se

NETWORKS IN BIOLOGY

Networks provide an intuitive and visually appealing way of organizing large amounts of data. They may convey a global view of a biological system, as well as its local details.

The general definition of a network is simple: It consists of a large number of *nodes*, which are unique elements (e.g. genes, proteins or metabolites) and *edges*, each of which corresponds to a pair-wise relationship between two of the nodes. The *relationship* that is encoded by the edge can have an extremely wide range of interpretations, making networks very flexible for the representation of biological knowledge.

The simple mathematical definition of networks leads to a very straightforward way to describe a biological network. In most cases, a single table (produced, e.g., in a spreadsheet program like Excel) is sufficient to encode all information, and these general network files can then be read by many analysis- and visualization tools. Such a table might look as follows:

1566-5240/06 $50.00+.00

First Node	Second Node	Edge Description
Transcription Factor X	Transcription Factor Y	up-regulates
Transcription Factor Y	Repressor Molecule Z	up-regulates
Repressor Molecule Z	Transcription Factor X	inhibits

This simple example would correspond to a small transcriptional network of a negative feedback loop. The description of the edges can be standardized, so that only certain ways of interaction are allowed, or it can contain additional quantitative information, such as the strength and kinetics of an up-regulation. Also, the nodes can be further defined by, e.g., quantitative data on the concentrations of the compounds in a particular physiological state. Such specifications, however, as well as potential layout information for the display of a network, will usually depend on the specific application software that is used, while the underlying general network description will be shared by almost any program.

There are two major types of network descriptions in biology. In the first type, an edge between two nodes implies a direct physical interaction between the nodes. An example is a protein interaction network. If two proteins are connected, they are known to bind to each other in some kind of biological assay. Other examples are metabolic networks, where metabolites are connected whenever they participate in the same enzymatic reaction or, alternatively, whenever one of them can be converted chemically into the other one. Such a *substrate* $\rightarrow$ *product* relationship indicates a further possibility of refining the definition of the network: All edges could have a defined *direction*, so that in the graphical representation they would be represented by arrows pointing from one node (source) to another (target).

A further example of a directed, physical network would be a gene regulation network, where an arrow from gene A to gene B would imply that the product of gene A influences the expression of gene B. An example is the fascinating model of sea urchin development by Davidson *et al*. [2] The automated inference of such regulatory networks from gene expression data is a major challenge in statistical bioinformatics, and a number of publications report successes in this area [3-6]. These are now being complemented by the results of genetical genomics studies, that combine expression measurements and genotyping in large populations to allow more precise causal inference than any of the two approach alone [7-9].

The second major type of network involves conceptual, rather than physical, interactions between nodes. For example, an edge could translate as

- "Gene A participates in the same physiological process as Gene B" (functional network), or
- "The expression changes of Gene A and Gene B look the same in many experiments" (expression correlation network; see Fig. **1** for an example), or
- "A combined knock-out of Gene A and Gene B is always lethal" (genetic interaction network), or even
- "Gene A is very often mentioned in the same papers as Gene B" (co-citation network).

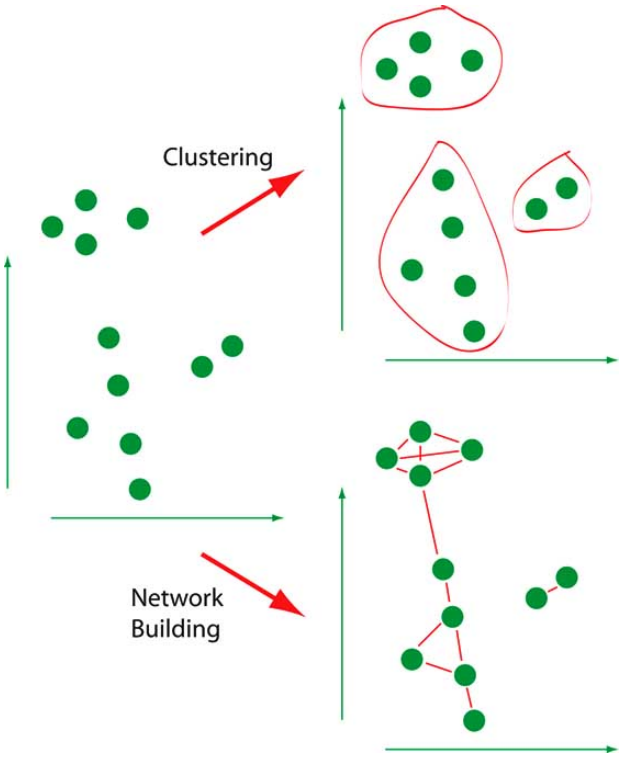

Fig. (1). Expression correlation networks. Gene expression patterns in multiple tissues and conditions can be visualized in multi-dimensional space (reduced to 2 dimensions in the figure). The closer two dots (genes or samples) are, the more similar are their expression patterns. In one highly popular approach (clustering), the aim is to identify groups of genes (or samples) that are similar and distinct from other samples. In the network approach presented in the present paper, genes are connected when their similarity (expression correlation) is above a certain threshold, i.e. if they are close in expression space. This may give more flexibility than "hard" clustering, but interpreting the results requires biological expertise.

All of these conceptual networks may be easier to generate than the physical networks. On the other hand, much of their usefulness derives from an inverse guilt-by-association principle: Genes that are connected by one of the more abstract concepts (e.g. by being mentioned in the same papers) are also expected to be biologically "close", i.e. they may

participate in the same physiological processes or influence the same complex disease phenotypes. A very active research area in current bioinformatics aims at generating “physical” networks (such as gene regulation pathways) from “conceptual” networks (such as co-expression information) [10].

NETWORK TOPOLOGY AND PHYSIOLOGICAL FUNCTION

One of the most striking findings of network analysis in biology has been that the overall structure of the networks is far from random. One could have assumed that the average probability that two genes are associated in one way or another is fairly uniform, but this is not the case. In general, all biological networks so far examined are characterized by a small number of highly connected nodes, while most other nodes have few connections. This is known as a power-law distribution of connections. The highly connected nodes act as hubs that mediate interactions between other nodes in the network [11] (see Fig. **2** for examples of network types). This has an important implication; hubs have key roles in the networks. A “hub-centered” design, which is similar to the layout of the internet, yields good overall stability of a network against malignant changes (“attacks”). This stability arises because the majority of nodes are low connectors which can be deleted without major global effects. However, hubs are particularly vulnerable, because changes to them will have drastic consequences all over the network. Evolution of a physiological system should lead to a compromise between increased vulnerability of hubs and increased stability of the complete system.

One of the first examples of a network theoretical analysis in biology was provided by Jeong *et al*. who showed that the connectivity in protein interaction networks in yeast cells followed an approximate power-law distribution [12]. Thus, a small number of proteins were highly connected to many other proteins. This property, which has become famous as the “power-law distribution of degrees” [11], has important consequences for the potential behavior of the system. For example, because of the sensitivity of hubs to mutation, any change in them will have large consequences for the entire system. Thus, they will often be represented by essential genes. In fact, deletion of “hub” proteins was associated with increased yeast cell lethality compared to deletion of less connected proteins [12]. This supports a key role for hubs. These properties of biological networks are also popularly referred to as "scale-free" topology, borrowing a term that was originally developed to characterize similar networks in theoretical physics. However, Khanin and Wit [13] present compelling evidence supporting the conclusion that experimental networks are not scale-free in the technical, mathematical sense. This has to be kept in mind when the evolution and relevance of the topological properties of biological networks are discussed. This is still an area of active and controversial research. Another interesting biological conclusion from network theory is that hub genes are highly conserved through evolution [11]. By contrast, mutations in other genes are more likely to be transmitted because they are less likely to affect survival.

The power-law distribution of degrees is closely related to the so-called 'small-world property' of biological networks [3], which means that every node can affect or associate to each other node *via* a small number of steps. This potentially has huge implications for complex disease traits: Small-world behavior of physiology and gene regulation raises the question if a change almost anywhere in the system may lead to pervasive pleiotropic effects. In a highly-connected, small-world neighborhood of the biological system, a change in any (albeit non-

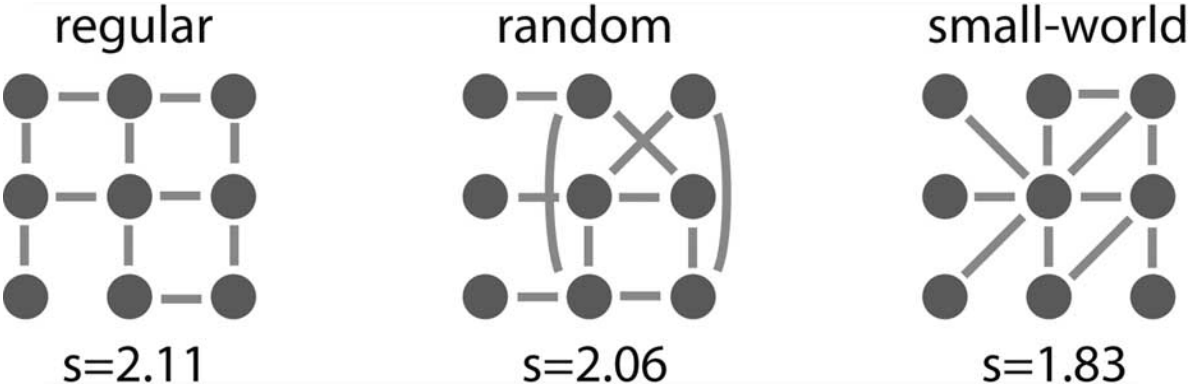

Fig. (2). Examples of networks. The figure shows three general types of network topology. In the first case, nodes are connected in a regular pattern, all nodes tending to have very similar connectivities (*regular network*). In the second case, nodes are randomly connected, any possible connection being as likely as any other (*random network*). In the third case, nodes are also randomly connected, but in such a way that one node, the hub, is much more strongly connected than would be expected in an entirely random network (*small-world network*). As the name implies, this leads to a reduction in the average path length, *s*, as indicated below the networks. This effect would be even stronger in larger networks. All three examples have 9 nodes and 11 edges.

essential) gene has the possibility to affect many other genes and functions, and, although these effects will often be mild and not always of pathophysiological relevance, they may still be picked up by microarrays and can contribute to the interpretation of the global physiological picture.

Biological networks are not only non-random in their global properties (such as their small-world characteristics), but also show important patterns at the local level. Biological entities show a strong inclination to be connected in characteristic ways. The shape and meaning of such "network motifs" [14] will vary widely between different types of network, but may be diagnostic for a certain kind of biological function. For example, in gene regulatory networks, a gene that is influencing the expression of many downstream genes may stand out as an important regulator, most likely a transcription factor or signaling molecule. A problem may be that many transcription factors are not transcriptionally regulated and therefore will not be picked up in classical microarray studies. Genetical genomics experiment may be able to overcome this important limitation. On the other hand, in a protein interaction network, a gene product that binds to a large number of other gene products, which in turn are connected to each other, can confidently be predicted to be a subunit of a protein complex. Considering that gene-gene regulatory connections can be either positive (activation) or negative (inhibition) even more interesting patterns emerge, for example an overrepresentation of certain feedback loops that can stabilize or amplify expression responses [14]. Genes that are involved in such network motifs can then be assigned to putative roles based purely on their connectivity, even without requiring additional biological information about their actual biochemistry.

NETWORKS AND MICROARRAYS

All these findings suggest that networks can be useful to organize large amounts of biological information, in particular gene expression data. Networks may for example, help to find modules of functionally related genes and "hub" genes with key regulatory roles. In this section we review applications of network approaches to microarray analysis.

Carter *et al*. used a very straightforward approach and built a gene correlation network from expression data, in which genes were connected when their expression profiles were correlated to a significant degree [15]. The authors report an apparently scale-free structure of these networks, but point out that the clustering coefficient is much larger than would be expected in a scale-free network. They interpret this as an indication of modularity in the network. The most likely explanation for such an effect is the presence of gene groups (modules) that are all regulated strongly (and exclusively) by a common transcription factor (or combination of transcription factors). An example of such a module, which is widely observed in many microarray studies, is the cluster of ribosomal genes. These respond strongly and in a coordinated fashion in gene expression studies in many organisms, cell types and conditions. Carter *et al*. [15] also observe an enrichment of lethal knock-out phenotypes among the high-degree nodes of their network. In addition they find an enrichment of condition-specific genes among the hubs. Both of these observations indicate that even in a simple correlation network like this, network theory can predict important physiological factors.

Many other microarray studies of different organisms and conditions have revealed conserved modules of co-expressed genes. Such modules correspond to distinct functions. In one such study based on more than 3000 microarray experiments conserved co-expression modules were associated with specific cellular functions such as lipid metabolism and cell cycle [16]. Such a module could be used to predict the function of an unknown gene by comparing it with neighboring genes in the module. Bergmann *et al.* also report modules of conserved interaction networks in different organisms [17].

In another study, Lee *et al*. [18] examine co-expression relationships of genes across 60 different large human data sets, comprising a total of 3924 microarrays. They identify genes that are reproducibly co-expressed in the various studies at high statistical significance and build a high-confidence network of 8805 genes connected by 220,649 co-expression links. This number is much higher than in the previous study, as only a single organism albeit in widely diverse conditions was considered. The authors report a clear preponderance of positive expression correlations, perhaps because the underlying molecular apparatus is designed to co-activate/co-repress genes, rather than changing them in opposite directions.

In agreement with Stuart *et al*. they detect functionally related clusters in the network that correspond to processes like protein biosynthesis and cell-cycle [16]. They also report that even co-expressions that are observed only in a few datasets are biologically informative and enriched for non-random relationships.

Comparative analysis can also be used to find conserved co-expression modules involved in cellular processes and disease. For example, McCarrol *et al.* found common modules that were associated with aging in worms and flies. These modules included orthologous genes involved in mitochondrial function, DNA repair and cellular transport. A similar approach was used to find modules involved in cancer. Sweet-Cordero *et al.* identified a module associated with lung cancer in three mouse models and a corresponding module of orthologous genes in human cancer [19]. The approach can also be extended to describe regulatory modules containing both regulatory factors and the corresponding down-

stream module of co-regulated genes. Such a module was identified in mice and its human orthologue is associated with cancer [20]. These experiments indicate how modules can be used for meaningful organization of microarray data on a global scale.

In the context of human microarray studies Basso *et al*. recently showed that a gene regulatory network reconstructed from B lymphocyte expression data had approximately scale-free properties [21]. They focused on the sub-network centered around the transcription factor myc. Using bioinformatics methods they found that several of the genes in the sub-network contained response elements of *myc*, many of which were verified experimentally. Hub genes were enriched among the *myc* targets, confirming its central (patho-)physiological role as a master regulator of cell proliferation [22].

About 5% of the B lymphocyte genes were hubs. It is interesting to note that about the same percentage of eukaryotic genes are transcription factors [23]. This could indicate that a large number of hubs are transcription factors. Taken together, this data supports once more the validity of applying network theory to networks derived from gene expression data from microarray studies.

Moreover, Segal *et al*. suggested that modules can be used as basic building blocks when attempting to understand biological effects [20]. Modules are easier to interpret than hundreds or thousands of differentially expressed genes. Also, subtle but coordinated changes of genes in modules are easier to discern than changes in individual genes. An early example of linking modules to biological effects was given by a study of apoptosis in diabetic muscle [24]. A specific module was linked to apoptosis. A meta-analysis of multiple microarray studies of cancer reported a large compendium of cancer-specific modules [20]. These modules contained genes that were co-expressed in many different experiments. The modules corresponded to different functions such as cell cycle or growth regulation. These modules could be represented in a causal chain, with transcription factors upstream and biological functions, such as cell growth or apoptosis, down-stream. Together, these modules formed a network map of transcriptional changes in cancer. The map or modules within it could be used to gain a mechanistic understanding of the disease process. For example, decreased expression of genes in a growth inhibitory module in acute leukemia, suggested a possible explanation for uncontrolled growth in this form of cancer. Other modules were shared across many different forms of cancer, indicating a common disease mechanism. Combinations of such modules could be associated with different types and stages of cancer. This showed the usefulness of analyzing complex processes in terms of modules, rather than individual genes.

NETWORKS IN COMPLEX DISEASE

Associating modules with transcription factors and down-stream cellular processes in causal chains, also shows how directionality can be introduced in network models. One way to change from associative to directed, causal networks was described by Chesler *et al*. [8]. Their work was based on the observation that quantitative differences in gene expression are heritable. They used a panel of several inbred generations of mice. Their aim was to link heritable differences in gene expression to genetic loci in these mice. In other word gene expression was treated as a quantitative trait and the loci identified as quantitative trait loci (QTL). mRNA samples from each inbred mouse line were analyzed with microarrays. It was shown that heritable differences in gene expression could be related to specific genetic regions (QTLs). By inference, these QTLs should contain either *cis*- or *trans*-regulatory elements that cause the expression difference. If a differentially expressed gene was located in a QTL responsible for its expression regulation, that gene was expected to be *cis*-regulated, e.g. by having a polymorphism in its upstream regulatory region. Trans-regulated modules, which are not themselves polymorphic, but respond to genetic differences in some regulatory gene, could contain more than a thousand genes. Another interesting finding was that, since phenotypic QTLs were also known for the multiple mouse generations, regulatory modules could also be linked to physical traits.

If network models can be built that describe transcription factors, modules, and down-stream biological effects such as cellular processes or disease, the next step in a top-down analysis would be to dissect the models to find individual genes with key functions.

In natural cellular processes transcription factors are key regulatory genes, since they are up-stream regulators of multiple genes. Indeed, the work by Chesler *et al*. suggested that a single transcription factor could regulate (directly or indirectly) the expression of thousands of genes [8]. Also, from a networks perspective transcription factors are likely to have key roles since they are often highly connected hubs. It remains to be demonstrated that transcription factors or other highly connected genes also have key roles in complex diseases.

Because hubs are more often essential for survival, deleterious mutations in them are less likely to be transmitted across generations than in the case of non-essential genes. Indeed, as stated above, hubs are more conserved than non-hubs [11]. By contrast, yeast cell experiments indicate that mutations in non-essential genes can accumulate in populations [25]. If these mutations occur in genes that do not interact, the risk of affecting cell fitness is small. However, yeast cell experiments involving mutations in pairs of genes indicated that mutations in pairs of interacting genes are much more likely

affect cell fitness. The authors reasoned that the same principle could be applied to complex diseases.

In other words, mutations in non-essential genes occur frequently as they are less likely to affect the evolutionary fitness of a carrier. However, if such mutations become frequent enough in the population, at one point it becomes likely that a single person will carry mutations in interacting non-essential genes. And while none of the individual mutations have serious effects, their combination may be severe and can result in disease. An example of such a human disease is retinitis pigmentosa, which is caused by mutations in two genes that are asymptomatic singly, but together cause the disease [26]. While this is not yet a "complex disease", it serves well to illustrate the general principle. Another related example is provided by cystic fibrosis. Although this is caused by mutations in a single gene, disease severity is strongly affected by mutations in several other genes [25].

It is of interest in this context, that the small-word property of networks implies that most genes may interact through only a small number of intermediary genes. This could increase the risk of disease-associated interactions between mutated genes. Linkage studies of complex diseases like allergy have described multiple loci, and it is possible that these contain mainly interacting non-essential genes. This possibility indicates how networks in combination with other sources of information such as linkage studies and bioinformatics methods could be used to pin-point combinations of mutations in different genes in complex disease. To our knowledge, this has not yet been shown.

FUTURE CHALLENGES

Apart from the quantitative complexity associated with analyzing 25,000 genes simultaneously, several biological factors may confound the interpretation of networks based on network models of transcriptomal changes:

- Changes in mRNA may not correspond to changes in protein levels due to variable translation or protein degradation.
- There may be post-translation modifications of proteins. For example, transcription factors are seldom regulated on the transcriptional level.
- Co-expressed genes may occur in different sub-cellular compartments and not interact physically in the same cellular process.
- Genetic polymorphisms may cause important changes which are not detected by microarrays, e.g. in posttranslational processing or protein stability.
- A vast majority of translational variation may be "physiological noise". Even though it has been shown that exact transcript levels will in many cases be under evolutionary control [27], some variation will be the spurious result of the high connectivity and small-world properties of transcriptional networks.
- Microarray analysis of complex tissues is confounded by variable cell populations, some of which may not be associated with the biological problem that is studied. Interestingly, computational methods comparing histological information with gene expression data may help address this problem [28].

CLINICAL APPLICATIONS

The examples described above suggest that identification of modules of functionally related genes may help to understand pathways involved in complex diseases as well as in response to treatment. This information could lead to identification of biomarkers for improved diagnosis as well as for individualized medication. DNA microarrays containing selected genes are already used diagnostically when treating patients with breast cancer [29]. It is conceivable that the same principles could be useful in other complex diseases. It should be noted, however, that at present there are no validated clinical tests based on applying network theory to microarray data.

SUMMARY

Organizing microarray data in network models may help understanding complex diseases on both the global and detailed scale. Although significant methodological and theoretical challenges remain, this may have clinical implications, such as identification of biomarkers and therapeutic targets.

REFERENCES

[1] Benson, M., Olsson, M., Rudemo, M., Wennergren, G. and Cardell, L. O. (2004) *Clin. Exp. Allergy,* **34**, 1001-1006.
[2] Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., Pan, Z., Schilstra, M. J., Clarke, P. J., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L. and Bolouri, H. (2002) *Science,* **295**, 1669-1678.
[3] Watts, D. J. and Strogatz, S. H. (1998) *Nature,* **393**, 440-442.
[4] Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003) *Nat. Biotechnol.,* **21**, 1337-1342.
[5] Qian, J., Lin, J., Luscombe, N. M., Yu, H. and Gerstein, M. (2003) *Bioinformatics,* **19**, 1917-1926.
[6] Wagner, A. (2004) *J. Comput. Biol.,* **11**, 53-60.
[7] Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B. and Friend, S. H. (2003) *Nature,* **422**, 297-302.
[8] Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., Threadgill, D. W., Manly, K. F. and Williams, R. W. (2005) *Nat. Genet.,* **37**, 233-242.

[9] Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004) *Nature,* **430**, 743-747.
[10] Wagner, A. (2001) *Bioinformatics,* **17**, 1183-1197.
[11] Barabasi, A. L. and Oltvai, Z. N. (2004) *Nat. Rev. Genet.,* **5**, 101-113.
[12] Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001) *Nature,* **411**, 41-42.
[13] Wit, E. and Khanin, R. (2006) *J. Comput. Biol.,* **13**(3), 810-818.
[14] Mangan, S. and Alon, U. (2003) *Proc. Natl. Acad. Sci. USA,* **100**, 11980-11985.
[15] Carter, S. L., Brechbuhler, C. M., Griffin, M. and Bond, A. T. (2004) *Bioinformatics,* **20**, 2242-2250.
[16] Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003) *Science,* **302**, 249-255.
[17] Bergmann, S., Ihmels, J. and Barkai, N. (2004) *PLoS Biol.,* **2**, E9.
[18] Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) *Genome Res.,* **14**, 1085-1094.
[19] Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R. and Jacks, T. (2005) *Nat. Genet.,* **37**, 48-55.
[20] Segal, E., Friedman, N., Kaminski, N., Regev, A. and Koller, D. (2005) *Nat. Genet.,* **37 Suppl**, S38-45.
[21] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005) *Nat. Genet.,* **37**, 382-390.
[22] Levens, D. (2002) *Proc. Natl. Acad. Sci. USA,* **99**, 5757-5759.
[23] van Steensel, B. (2005) *Nat. Genet.,* **37 Suppl**, S18-24.
[24] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003) *Nat. Genet.,* **34**, 267-273.
[25] Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Menard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A. M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H. and Boone, C. (2004) *Science,* **303**, 808-813.
[26] Kajiwara, K., Berson, E. L. and Dryja, T. P. (1994) *Science,* **264**, 1604-1608.
[27] Wagner, A. (2005) *Mol. Biol. Evol.,* **22**, 1365-1374.
[28] Stuart, R. O., Wachsman, W., Berry, C. C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M., Wang, Y., Sawyers, A., Kalcheva, I., Tarin, D. and Mercola, D. (2004) *Proc. Natl. Acad. Sci. USA,* **101**, 615-620.
[29] Garber, K. (2004) *Science,* **303**, 1754-1755.

Received: January 12, 2006 **Revised: April 25, 2006** **Accepted: April 26, 2006**